

Figure 4. Same as Figure 3, except that the means of the variables associated with all components past component one were now also chosen uniformly randomly, though from between 1.9 and 2.1.

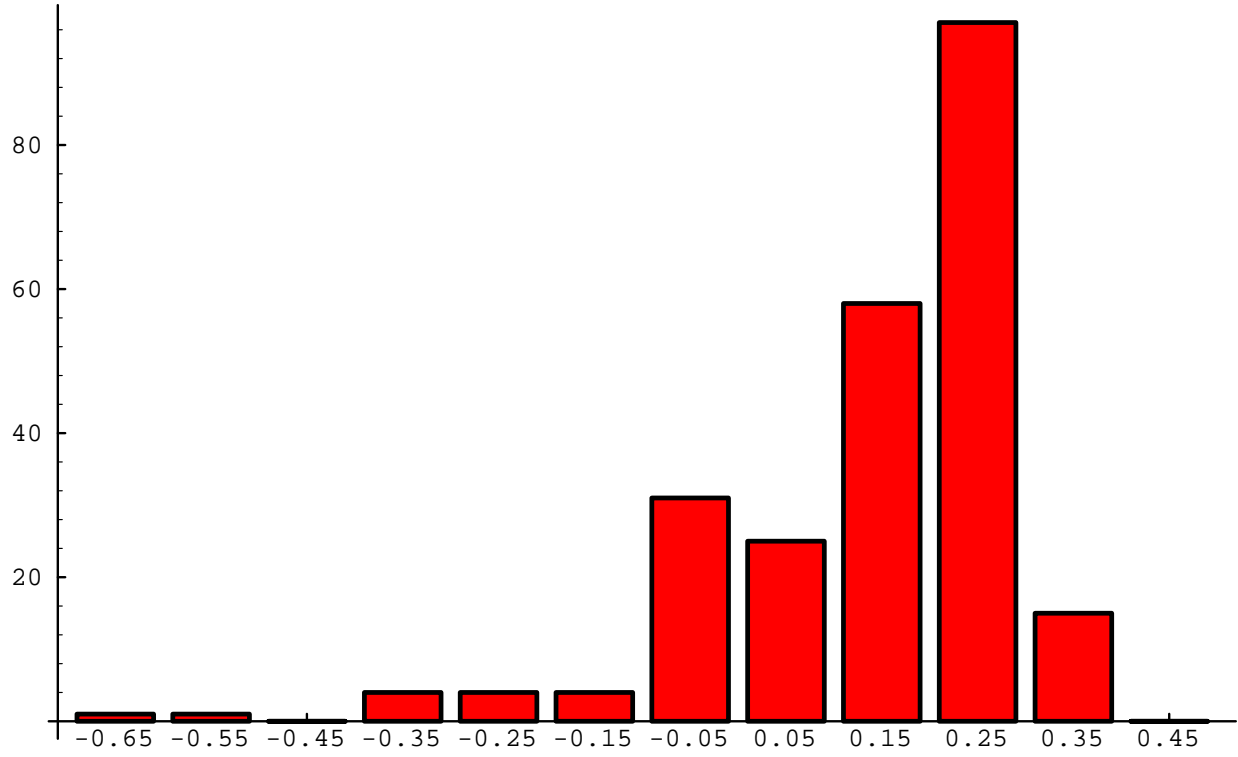


Figure 3. Histogram of the differences in ρ between $\epsilon(\vec{b})$ and $\epsilon(\vec{c})$ for the same situation as in Figure 2, except N is increased to 10, 240 experiments were conducted, $\vec{c} = \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1\}$, and all components of $\vec{b} < .25$ were zeroed out. The means for component one for both actions were chosen uniformly randomly between 0 and 10.0, and the variances were chosen uniformly randomly between 0.0 and 3.0. The means of the other variables were all 2, for both actions, and the associated variances were chosen uniformly randomly between 0 and 5.0.

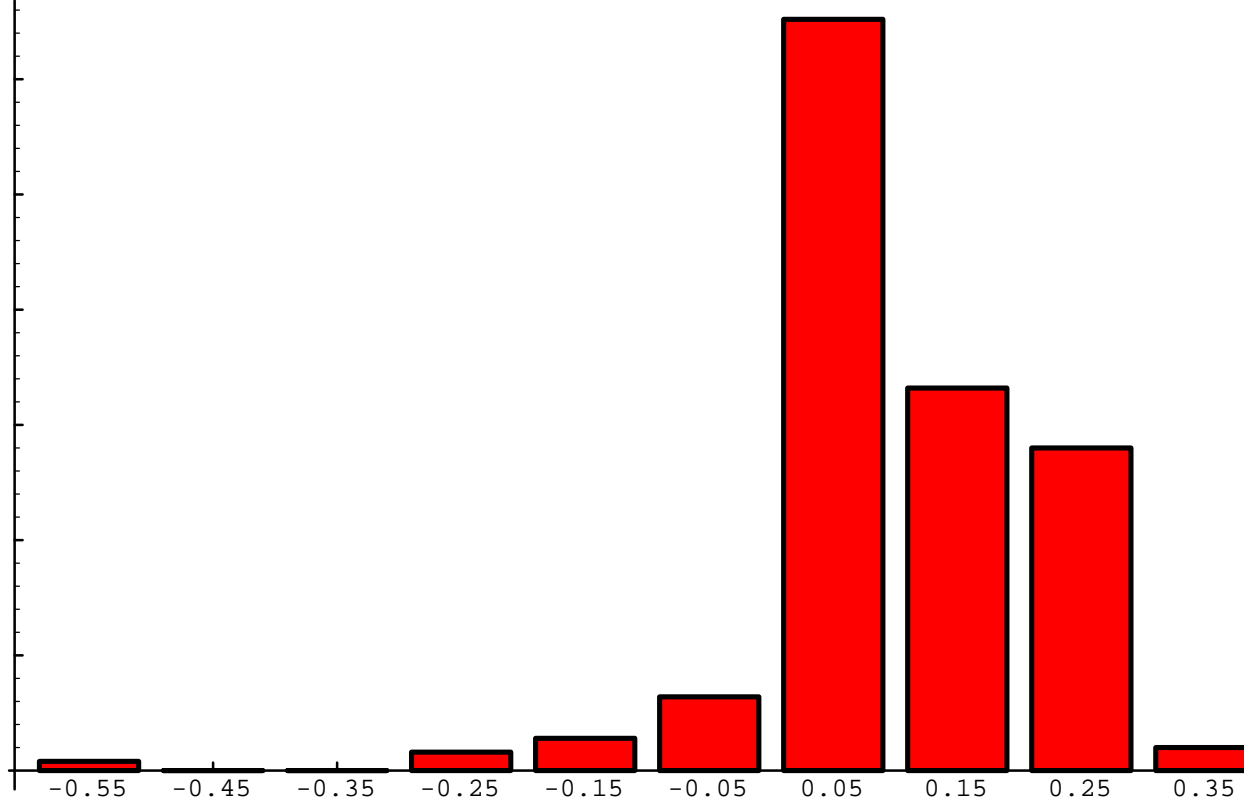


Figure 2. Histogram of the difference between ρ when $\vec{b} = \vec{c}$ and ρ for the teacher's \vec{b} . The total number of experiments was 350, M was 50, $m = 1$, $\vec{c} = (1, 1)$, and diagonal covariance matrices were used. The two variances for the first component of \vec{y} (one variance for each action) were both chosen by sampling the uniform distribution extending from 0.0 to 1.0, and for the second component by sampling the uniform distribution extending from 0.0 to 100.0. The components of $\vec{\mu}_1$ were chosen by randomly sampling a uniform distribution from 0.0 to 10.0. Both components of $\vec{\mu}_2$ were 2.

FIGURES.

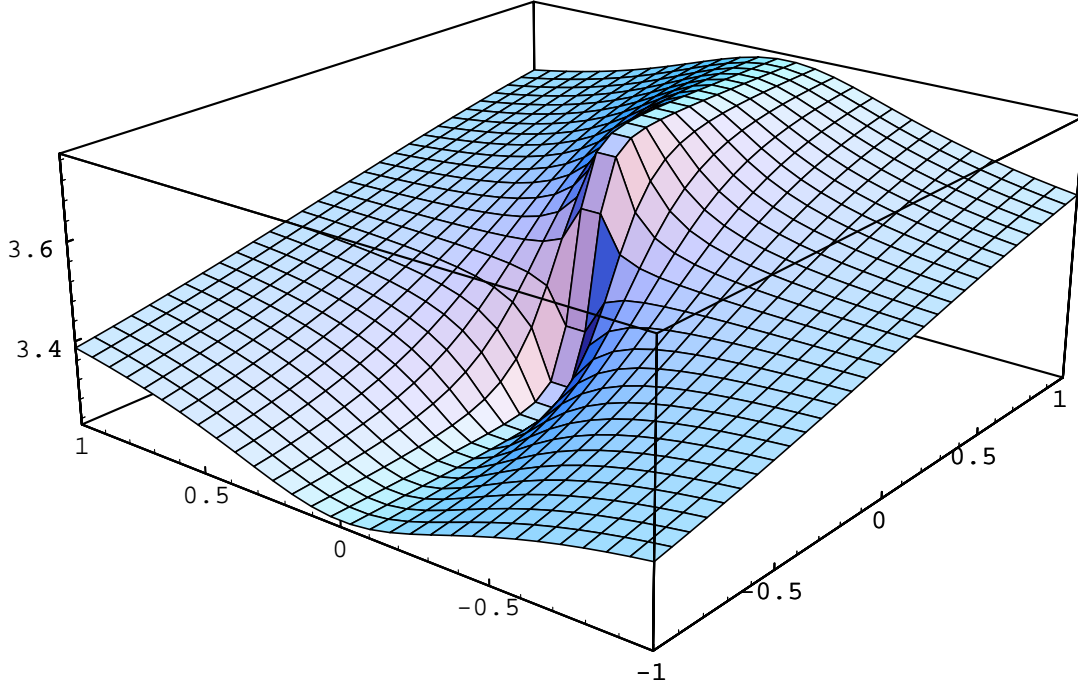


Figure 1. A plot of $\varepsilon(\vec{b})$ for $K = 2$, $N = 2$, $\tilde{D} = \tilde{0}$, $m = 1$, $\vec{c} = (1, 1)$, and diagonal covariance matrices. For action 1, $\vec{\mu}_1 = (1, 3)$, and $\vec{\mu}_2 = (0, 3)$. The two variances for both actions were 1 and 25 (one variance for each component of \vec{y}). The optimal \vec{b} is proportional to $(1, 0)$. For this \vec{b} , ρ was .16. In contrast, ρ for $\vec{b} = \vec{c}$ was .34; performance improved by using the optimal \vec{b} by over a factor of 2.

6. Jennings, N. R., Sycara, K. and Wooldridge, M., "A Roadmap of Agent Research and Development", *Autonomous Agents and Multi-Agent Systems*, **1**, 7-38, 1998.
7. Kaelbling, L. P., Littman, M. L. and Moore, A. W., "Reinforcement Learning: A Survey", *Journal of Artificial Intelligence Research*, **4**, 237-285, 1996.
8. Sutton, R.S., "Learning to Predict by the methods of temporal differences ", *Machine Learning*, **3**, 9-44, 1998.
9. Sutton, R. S. and Barto, A. G., "Reinforcement Learning: An Introduction", MIT Press, Cambridge, MA, 1998.
10. Sycara, K. "Multiagent Systems", *AI Magazine*, **19**, 79-92, 1998.
11. Watkins, C., and Dayan, P., "Q-Learning", *Machines Learning*, **8**, 279-292, 1992.
12. Wolpert, D. Tumer, K., and Frank, J. "Using Collective Intelligence to Route Internet Traffic", in *Advances in Neural Information Processing Systems - 11*, MIT Press, in press.
13. Wolpert, D., Wheeler, K., and Tumer, K., "General Principles of Learning-based Multi-Agent Systems", in *Proceedings of the Third International Conference of Autonomous Agents*, in press.
14. Wolpert, D., and Tumer, K., "A Survey of Collective Intelligence", in *Handbook of Agent technology*, J. M. Bradshaw (Ed.), AAAI Press/MIT Press, in press.
15. Wolpert, D., "On Bias Plus Variance", *Machine Learning*, **9**, 1211-1244, 1998.

$$= \frac{\int [\tilde{u} \cdot (\tilde{B}^T \tilde{D} \tilde{B}) \cdot \tilde{u}] e^{-\tilde{u} \cdot \tilde{E}^{-1} \cdot \tilde{u} / 2} d\tilde{u}}{(2\pi)^{N/2} \sqrt{\det(\tilde{E})}}, \text{ since } \tilde{E}^{-1} \text{ is diagonal and terms like } \tilde{u} \cdot (\tilde{B}^T \tilde{D}) \cdot \tilde{u}_1 \text{ have odd symmetry.}$$

Again using \tilde{E}^{-1} 's being diagonal and symmetry arguments, we can write our integral, getting

$$\text{C.3) } \sum_i \left[\frac{\int [u_i^2 (\tilde{B}^T \tilde{D} \tilde{B})_{ii}] e^{-u_i^2 \tilde{E}_{ii}^{-1} / 2} du_i}{2\pi \sqrt{\tilde{E}_{ii}}} \right] = \sum_i \tilde{E}_{ii} (\tilde{B}^T \tilde{D} \tilde{B})_{ii} = \text{Tr}(\tilde{B}^T \tilde{D} \tilde{B} \tilde{E}), \text{ since } \tilde{E} \text{ is diagonal. But } \text{Tr}(\tilde{B}^T \tilde{D} \tilde{B} \tilde{E}) = \text{Tr}(\tilde{D} \tilde{B} \tilde{E} \tilde{B}^T) = \text{Tr}(\tilde{D} \tilde{C}), \text{ by definition of } \tilde{B} \text{ and } \tilde{E}.$$

A similar result holds for \hat{G}_2 . Now use Thm. 2 to write $E(G \mid g_1, g_2, m) = \frac{(\hat{G}_1 + \hat{G}_2)}{2} + (\hat{G}_2 + \hat{G}_1) \left(\frac{1}{2} - C[E(\delta_m), \sigma^2(\delta_m)] \right)$. Plugging in gives the result claimed. **QED.**

REFERENCES

1. Bass, T., "Road to Ruin", *Discover*, 56-61, May, 1992.
2. Boyan, J. and Littman, M., "Packet Routing in Dynamically Changing Networks: A Reinforcement Learning Approach", in *Advances in Neural Information Processing Systems - 6*, J. Cowan and G. Tesauro and J. Alspector (Eds), 671-678, Morgan Kaufmann, 1994.
3. Claus, C. and Boutilier, C., "The Dynamics of Reinforcement Learning Cooperative Multiagent Systems", in *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, 746-752}, June, 1998.
4. Hardin, G., "The Tragedy of the Commons", *Science*, **162**, 1243-1248, 1968.
5. Helbing, D. and Treiber, M., "Jams, Waves, and Clusters", *Science*, **282**, 200-201, December, 1998.

$2p_L (\mu - \mu_L)^2$, which takes on its minimal value of $2p_L \mu^2$ when $\mu_L = 0$. So $p_L = \sigma^2 / 2\mu^2$, which is what one would expect from Chebychev's inequality.

In addition to this bound, since x is symmetric about μ , we also know that $p_L \leq 1/2$. This establishes (ii). **QED.**

APPENDIX C - Proof of Theorem 4.

Examining the terms in Thm. 2, we see immediately that $\hat{g}_1 = \vec{b} \cdot \vec{\mu}_1$ and $\hat{g}_2 = \vec{b} \cdot \vec{\mu}_2$. We can also immediately write $\hat{\sigma}_1^2 = E([\vec{b} \cdot (\vec{y}_1 - \vec{\mu}_1)]^2) = E(\vec{b}^T [(\vec{y}_1 - \vec{\mu}_1)^T (\vec{y}_1 - \vec{\mu}_1)] \vec{b}) = \vec{b}^T \cdot E((\vec{y}_1 - \vec{\mu}_1)^T (\vec{y}_1 - \vec{\mu}_1)) \cdot \vec{b} = \vec{b} \cdot \tilde{C}_1 \cdot \vec{b}$. Similarly, $\hat{\sigma}_2^2 = \vec{b} \cdot \tilde{C}_2 \cdot \vec{b}$. So $C(E(\delta_m), \sigma^2(\delta_m))$ is the cumulative distribution function of a Gaussian with mean $\vec{b} \cdot (\vec{\mu}_2 - \vec{\mu}_1)$ and variance $\frac{\vec{b} \cdot (\tilde{C}_2 + \tilde{C}_1) \cdot \vec{b}}{m}$, evaluated at 0. This is just $\{1 - \text{erf}\left(\vec{b} \cdot (\vec{\mu}_2 - \vec{\mu}_1) \sqrt{\frac{m}{2[\vec{b} \cdot (\tilde{C}_2 + \tilde{C}_1) \cdot \vec{b}]}}\right) / 2\}$.

The remaining terms to calculate are \hat{G}_1 and \hat{G}_2 . Writing it out,

$$\begin{aligned} \text{C.1) } \hat{G}_1 &= E(G \mid A = 1) = \int G(\vec{y}_1) P(\vec{y}_1) d\vec{y}_1 \\ &= \frac{\int (\vec{c} \cdot \vec{y}_1 + \vec{y}_1 \cdot \tilde{D} \cdot \vec{y}_1) e^{-(\vec{y}_1 - \vec{\mu}_1) \cdot \tilde{C}_1^{-1} \cdot (\vec{y}_1 - \vec{\mu}_1) / 2} d\vec{y}_1}{(2\pi)^{N/2} \sqrt{\det(\tilde{C}_1)}} \\ &= \vec{c} \cdot \vec{\mu}_1 + \vec{\mu}_1 \cdot \tilde{D} \cdot \vec{\mu}_1 + \frac{\int [\vec{y}_1 \cdot \tilde{D} \cdot \vec{y}_1 + \vec{y}_1 \cdot \tilde{D} \cdot \vec{\mu}_1 + \vec{\mu}_1 \cdot \tilde{D} \cdot \vec{y}_1] e^{-\vec{y}_1 \cdot \tilde{C}_1^{-1} \cdot \vec{y}_1 / 2} d\vec{y}_1}{(2\pi)^{N/2} \sqrt{\det(\tilde{C}_1)}} \end{aligned}$$

Now make the variable transformation $\vec{y} = \tilde{B} \vec{u}$ where $\tilde{B}^T \tilde{C}_1^{-1} \tilde{B} = \tilde{E}^{-1}$ for some diagonal matrix \tilde{E} , i.e., \tilde{B} diagonalize \tilde{C}_1^{-1} . Then $\det(\tilde{B}) = \sqrt{\frac{\det(\tilde{C}_1)}{\det(\tilde{E})}}$, and our integral becomes

$$\text{C.2) } \frac{\int [\vec{u} \cdot (\tilde{B}^T \tilde{D} \tilde{B}) \cdot \vec{u} + \vec{u} \cdot (\tilde{B}^T \tilde{D}) \cdot \vec{\mu}_1 + \vec{\mu}_1 \cdot (\tilde{D} \tilde{B}) \cdot \vec{u}] e^{-\vec{u} \cdot \tilde{E}^{-1} \cdot \vec{u} / 2} d\vec{u}}{(2\pi)^{N/2} \sqrt{\det(\tilde{E})}}$$

Now $\sum_{i=1}^K \int P_i(u) [\prod_{j \neq i} C_j(u)] du = \int \frac{d}{du} [\prod_j C_j(u)] du = 1$ always, by the chain rule. Accordingly, if we add and subtract $\hat{G}_K \sum_{i=1}^K \int P_i(u) [\prod_{j \neq i} C_j(u)] du$ from the last expression in Eq. (B.2), we get $\hat{G}_K + \sum_{i=1}^{K-1} (\hat{G}_i - \hat{G}_K) \int P_i(u) [\prod_{j \neq i} C_j(u)] du$. Plugging this into the definition of $\rho_{\{g_i\}}(K, m)$ gives the result claimed. **QED.**

APPENDIX B - Proof of Lemma 1.

Consider the case where $P(x)$ is not constrained to be symmetric about μ . The bound in (i) trivially holds for $\mu = 0$. Without loss of generality take $\mu > 0$. Define $p_R \equiv \int_0^\infty dx P(x)$ and p_L similarly. Define μ_R as the expectation of x restricted to the positive axis, $\int_0^\infty dx xP(x) / p_R$, and define μ_L similarly. Then by Jensen's inequality, for any fixed form of the distribution $P(x)$ over the $x < 0$, if we replace $P(x)$ for the $x > 0$ with $p_R \delta(x - \mu_R)$ we will decrease the variance of $P(x)$ over all of x , while not changing its mean. Therefore for a fixed μ_R, μ_L and p_R , the associated $P(x)$ having the minimal variance is $P(x) = p_R \delta(x - \mu_R) + (1 - p_L) \delta(x - \mu_L)$.

That minimal variance is $(\mu_R - \mu_L)^2 (p_R - p_R^2)$. The associated value of μ is just $p_R(\mu_R - \mu_L) + \mu_L$. Therefore $(\mu_R - \mu_L) = (\mu - \mu_L) / p_R$. Plugging into the formula for the minimal variance, we get $\sigma^2 = (\mu - \mu_L)^2 (p_R - p_R^2) / p_R^2$. Since μ_L must be ≤ 0 by definition and $\mu \geq 0$ by hypothesis, this minimal variance is minimized by having $\mu_L = 0$: $\sigma^2 = \mu^2 p_L / (1 - p_L)$.

This lower bound on the variance is monotonically increasing as a function of p_L . Accordingly, this same formula gives us an upper bound on the p_L that are compatible with a given variance. Inverting, we see that that bound is just $\sigma^2 / (\sigma^2 + \mu^2)$. This establishes (i).

To establish (ii), we start the same way, replacing $P(x < 0)$ with $p_L \delta(x - \mu)$. Since x is symmetrically distributed about μ , we must concurrently replace $P(x > 2\mu)$ with $p_L \delta(x - (\mu - \mu_L))$. As our next step, we replace $P(0 \leq x \leq 2\mu)$ with $(1 - 2p_L) \delta(x - \mu)$. Doing all this leaves $E(x)$ and $P(x < 0)$ unchanged, and also leaves x symmetric about μ , while decreasing σ^2 . This minimal variance is

APPENDIX A - Proof of Theorem 1.

Writing it out,

$$\text{B.1)} \quad P(\text{naive student chooses action } i \mid K, m, \{g_i\}) =$$

$$\int \Theta(s_i - \max[s_1, \dots, s_{i-1}, s_{i+1}, \dots, s_K]) \prod_{j=1}^K P_{j, g_j, m}(s_j) ds_1 \dots ds_K.$$

To evaluate this, it helps to consider the density function $P_{-i}(u)$, defined as the density over the random variable $\max[g_1(\hat{y}_1), \dots, g_{i-1}(\hat{y}_{i-1}), g_{i+1}(\hat{y}_{i+1}), \dots, g_K(\hat{y}_K)]$, evaluated at the value u . This is because by change of variables, our integral can be rewritten as $\int \Theta(s - u) P_{i, g_i, m}(s) P_{-i}(u) ds du$. Now by definition, $P_{-i}(u) = \frac{d}{du}(\Pr[\sum_t g_1(\hat{y}_1(t)) \leq mu, \dots, \sum_t g_{i-1}(\hat{y}_{i-1}(t)) \leq mu, \sum_t g_{i+1}(\hat{y}_{i+1}(t)) \leq mu, \dots, \sum_t g_K(\hat{y}_K(t)) \leq t])$, where $\Pr(x)$ is defined as the probability of event x . Since all the events in the argument of the probability in our expression are independent, we can rewrite this as $P_{-i}(u) = \frac{d}{du}(C_1(u) \times \dots \times C_{i-1}(u) \times C_{i+1}(u) \times C_K(u))$ where for shorthand we're defining $C_i(u) \equiv C_{i, g_i, m}(u)$.

Plugging this in, we get $\int \Theta(s - u) P_{i, g_i, m}(s) \frac{d}{du}[\prod_{j \neq i} C_j(u)] ds du$ as our integral. Performing the inner integral over s gives $1 - C_i(u)$. Plugging this into Eq. (B.1) gives

$$\text{B.2)} \quad E(G \mid \text{naive student}, K, m, \{g_i\})$$

$$= \sum_{i=1}^K \hat{G}_i \int [1 - C_i(u)] \frac{d}{du}[\prod_{j \neq i} C_j(u)] du$$

$$= \sum_{i=1}^K \hat{G}_i \left[1 - \int C_i(u) \frac{d}{du}[\prod_{j \neq i} C_j(u)] du \right]$$

$$= \sum_{i=1}^K \hat{G}_i \int P_i(u) [\prod_{j \neq i} C_j(u)] du \quad (\text{after integrating by parts}).$$

2. Throughout this paper we will assume that there are no singularities in any of our distributions and none of the $\{g_i(\cdot)\}$ have “plateaus” of nonzero measure across which they take on a constant value. Accordingly, this argmax is unique, with probability 1.

3. In practice, it is often useful to evaluate $P_{i,f,m}(x)$ by multiplying m times the inverse Fourier transform of $[F(P_{i,f,1}(x))(k)]^m$ evaluated at mt , where $F(P_{i,f,1}(x))(k)$ is the Fourier transform of $P_{i,f,1}(x)$, evaluated at k .

4. The canonical example of such a case would be where the prior over μ_1 and μ_2 is uniform up to very large cutoffs. For such a case the posterior of ξ is $\int d\mu_1 d\mu_2 \delta(\xi - (\mu_2 - \mu_1)) P(\mu_1, \mu_2 | \{\hat{y}_1(t), \hat{y}_2(t)\}) \propto \int d\mu_1 d\mu_2 \delta(\xi - (\mu_2 - \mu_1)) P(\{\hat{y}_1(t)\} | \mu_1) P(\{\hat{y}_2(t)\} | \mu_2)$. Up to overall normalization constants, this integral is just the distribution of the difference of two random variables μ_1 and μ_2 , distributed according to $P(\{\hat{y}_1(t)\} | \mu_1)$ and $P(\{\hat{y}_2(t)\} | \mu_2)$, respectively. Since those two distributions are just Gaussians, we see that the posterior is just a Gaussian over ξ , having mean $[\sum_{1 \leq t \leq M} (\hat{y}_2(t) - \hat{y}_1(t))] / M$, and having variance σ_i^2 / M for each component i if the covariances of \hat{y}_1 and \hat{y}_2 are diagonal. That mean is, by definition, $\bar{\xi}$. In turn, as discussed above just before Thm. 2, since each σ_i^2 is the variance of the difference of two random variables (namely the i 'th component of \hat{y}_2 and the i 'th component of \hat{y}_1), it equals the sum of the variances of those two variables. Finally, those two variances can be estimated from the data directly, for example in the same calculation that estimates the two \tilde{C}_i , say by using maximum likelihood. This provides us with our variances: $\bar{\sigma}_i^2 = \frac{\sigma_i^2}{M}$.

where rather than simply setting the reward signal of a single student to optimize its utility, there are multiple students, and one must set their reward signals so that their collective behavior optimizes a global utility. What makes this problem so challenging is that in addition to addressing the “optimal teaching” kinds of issues investigated in this paper, in choosing each of the student’s reward signals we must also ensure that those signals induce the students to work cooperatively as far as the global goal is concerned, rather than at cross-purposes. In particular, we must ensure that the system does not exhibit tragedy of the commons phenomena [4], like traffic jams and bottlenecks [1, 5].

This variant of this paper’s topic is known as “Collective Intelligence”. Preliminary work on collective intelligence, including an overview, and applications to network routing, the El Farol Bar problem, and the leader-follower problem, can be found in [14, 12, 13], respectively. In addition to extending the results of this paper to more complicated learning scenarios and students, future work also involves incorporating these results into the collective intelligence domain.

FOOTNOTES

1. Of course, one can always dispute the validity of any particular choice of estimator, this one included. Our purpose in this paper is not to engage in (potentially endless) disputes about what estimator the student should use. For *any* choice, there will be an associated calculation we can perform of how best to distort the reward signal the student receives. In general, that optimal distortion will be non-zero. This paper is simply the investigation of this issue of how to distort the reward signal for one estimator, an estimator that is both very reasonable *a priori* and imposes an extremely small computational burden on the student. The latter point is especially important when one is concerned with massive MAS’s, many of the agents in which are computationally weak. See [14].

smallest of the components of the filter set to 0 after the ascent has completed. With this approximation, the communication and computational overheads in generating the student’s reward signal at each moment in the teaching phase is minimal. We also intend to investigate “parallelizing the teacher”, by distributing to the computational devices associated with each random variable an approximate version of the calculation of whether the associated component of the linear filter is low enough to be set to zero. This minimizes both the computational and communication burdens on the teacher, compared to having the teacher receive and process all the data from all the random variables.

Other future work involves calculating the optimal functions $\{g_i(\cdot)\}$ for the case of the Bayesian calculation with a Gaussian approximation to the posterior. This contrasts with the calculation presented above as our “special case”, which is of the optimal $g(\cdot)$ that is a linear function of its argument and is independent of A . Of particular interest is the case where the prior over the vector $\{E(\hat{y}_i)\}$ is biased towards having many components be independent of i , since that should be the case in large MAS’s, where many random variables in the environment don’t depend on the student’s actions. Also of interest is using tractable priors over the covariances of the $\{\hat{y}_i\}$. Regardless of the priors one uses, one practical concern with using this kind of more general $\{g_i(\cdot)\}$ is that, having more degrees of freedom than the $g(\cdot)$ calculated here, it may be prone to overfitting the data, especially if not all of our distributions are Gaussians.

Other future work involves investigating schemes for distorting reward functions in more complicated RL scenarios than the one considered in this paper. Such work would consider scenarios in which utility is not an undiscounted sum of rewards each of which only depends on a single action by the agent. In particular, such work would consider alternatives to the usual Q-learning and TD types of schemes in which the utility function is distorted “with malice aforethought” to improve the performance of the RL algorithm. Such distortions could potentially be used to address issues like credit assignment, the exploration-exploitation trade-off, etc., in addition to the signal/noise issues explored in this paper.

An extraordinarily rich and challenging variant of the work in this paper concerns situations

the associated variances were chosen uniformly randomly between 0 and 5.0.

A total of 240 experiments were conducted. When the threshold for zeroing a component of \vec{b} was 0 (i.e., no components were zeroed), the difference in ρ between $\epsilon(\vec{b})$ and $\epsilon(\vec{c})$ was $.137 \pm .025$. When we zeroed out all components of \vec{b} that were less than .25, the average difference in ρ between $\epsilon(\vec{b})$ and $\epsilon(\vec{c})$ was $.142 \pm .027$. The histogram of those differences in ρ for this second case are presented in Figure 3. On average, 76% of the components of \vec{b} were zeroed out. In 15 of the 240 experiments the first component of \vec{b} was (erroneously) zeroed out.

We then conducted a second set of zeroing out experiments identical to these first ones, except that the means of the variables associated with all components past component one were now also chosen uniformly randomly, though from between 1.9 and 2.1. Without zeroing the average difference in ρ between $\epsilon(\vec{b})$ and $\epsilon(\vec{c})$ was $.150 \pm .025$. When all components with values less than .25 were zeroed out, the difference in ρ between $\epsilon(\vec{b})$ and $\epsilon(\vec{c})$ was $.156 \pm .027$. The histogram of those differences in ρ for this second set of zeroing- \vec{b} 's experiments are presented in Figure 4. On average, 79% of the components of \vec{b} were zeroed out. In 9 of the 240 experiments the first component of \vec{b} was (erroneously) zeroed out.

Clearly for this problem at least, zeroing out small-enough components of \vec{b} results in no degradation in performance.

CONCLUSIONS

This paper demonstrates that distorting the reward function can result in major improvements in performance of a reinforcement learning algorithm, both in theory and in simulations. In the future we plan to extend our investigation in many respects. One is to consider the setting of linear filter reward functions via gradient ascent over the kinds of model-independent, student-independent approximations to the surface of posterior expected true utility that were discussed after Thm. 4. In particular, we plan to investigate approximating such a gradient ascent by having the

teacher with an observation phase. The teacher then used the data collected during that phase to set the parameters of the approximation to the posterior expected reward presented at the end of Section 4. It then ran a gradient ascent on that surface to find an optimal \vec{b} . $\epsilon(\vec{b})$ was then calculated using the actual means and variances of the \vec{y} , giving the expected performance of a student when using that \vec{b} to set its reward signals during a subsequent teaching phase. The value of ρ for these rewards was compared to that of $\epsilon(\vec{c})$ to get a final quantification of how much the Bayesian teacher managed to benefit the student.

The result of these experiments is presented in Figure 2 as a histogram of the difference between ρ when $\vec{b} = \vec{c}$ and ρ for the teacher's \vec{b} . The total number of experiments was 350, and M was 50. $m = 1$, $\vec{c} = (1, 1)$, and diagonal covariance matrices were used. The two variances for the first component of \vec{y} (one variance for each action) were both chosen by sampling the uniform distribution extending from 0.0 to 1.0. The two variances for the second component were both chosen by sampling the uniform distribution extending from 0.0 to 100.0. The components of μ_1 were chosen by randomly sampling uniform distributions, and similarly for the components of μ_2 . Those distributions were both Dirac delta functions about the same value for component 2, centered about 2. For component 1, for both actions, the upper bound of the distribution was 10, and the lower bound of the distribution was 0. The difference in ρ extended from a low of -.53 to a high of .34. The average was .098, +/- .013. A total of 320 out of the 350 experiments resulted in a positive difference in the ρ 's. Clearly the Bayesian teacher provides very significant benefit to the student.

Finally, we have conducted some preliminary investigations of our scheme for setting some of the components of the Bayesian teacher's \vec{b} to 0. We increased N to 10, while keeping diagonal covariance matrices, $M = 50$ and $m = 1$. $\vec{c} = \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1\}$. As in the experiments that resulted in Figure 2, we still only had component one matter, i.e., for all other components the means were the same for both actions. The means for component one for both actions were chosen uniformly randomly between 0 and 10.0, and the standard deviations were chosen uniformly randomly between 0.0 and 3.0. The means of the other variables were all 2, for both actions, and

$$\varepsilon(\vec{b}, \vec{z}) \approx \varepsilon(\vec{b}, \vec{z}) +$$

$$\sum_{1 \leq i \leq N} \left[\frac{\bar{\sigma}_i^2}{2} \right] \left[\sqrt{\frac{M}{2\pi |\vec{b} \cdot (\tilde{C}_2 + \tilde{C}_1) \cdot \vec{b}|}} \right] \left[2b_i b_i - \frac{M b_i b_i (\vec{b} \cdot \vec{z})(\vec{c} \cdot \vec{z})}{|\vec{b} \cdot (\tilde{C}_2 + \tilde{C}_1) \cdot \vec{b}|} \right] e^{-\left(\frac{(\vec{b} \cdot \vec{z})^2}{|\vec{b} \cdot (\tilde{C}_2 + \tilde{C}_1) \cdot \vec{b}|} \right) \left(\frac{M}{2} \right)}.$$

This approximation has the following reasonable properties:

- 1) It is proportional to the magnitude of \vec{c} ;
- 2) It is invariant under rescaling of \vec{b} ;
- 3) $(\bar{\sigma}_i^2 = 0 \ \forall i) \Rightarrow$ only the ε term contributes;
- 4) Changing \vec{b} to increase $\vec{b} \cdot \vec{c}$ while keeping everything else fixed is good, in general.

5. EXPERIMENTS

To consider the special case of Sections 3 and 4 in more detail, Figure 1 presents the function for the case of $K = 2$, $N = 2$, $\tilde{D} = \tilde{0}$, $m = 1$, $\vec{c} = (1, 1)$, and diagonal covariance matrices. For action 1, $\vec{\mu}_1 = (1, 3)$, and $\vec{\mu}_2 = (0, 3)$. Note that these means are identical for component 2 of \vec{y} - that component serves purely as noise. The two variances for action 1 were 1 and 25 (one variance for each component of \vec{y}). The two variances for action 2 were 1 and 25. (A large variance for a component 2 that contributes only noise, with $K = 2$, is equivalent to small variances for many components all of which contribute only noise, with $K > 2$.) The optimal \vec{b} is proportional to $(1, 0)$. For this \vec{b} , ρ was .16. In contrast, ρ for $\vec{b} = \vec{c}$ was .34. Both ρ 's are less than .5, in agreement with Thm. 3(ii). Performance improved by using the optimal \vec{b} by over a factor of 2.

To test how much of this potential improvement can be actually realized by our Bayesian teacher, we ran a set of computer simulations. In each of these the means and variances of the \vec{y}_i were randomly chosen, and the resultant distributions were sampled M times to provide the

4. THE POSTERIOR OPTIMAL TEACHER FOR OUR SPECIAL CASE

In the real world the teacher doesn't know μ_1 , μ_2 , \tilde{C}_1 , or \tilde{C}_2 , but must estimate them from the data acquired during the teacher's observation phase. Accordingly, the teacher's task is to choose the \vec{b} that maximizes the posterior expected value of G , $E(G \mid \vec{b}, m, \text{data}) = \int E(G \mid \vec{b}, \mu_1, \mu_2, \tilde{C}_1, \tilde{C}_2) P(\mu_1, \mu_2, \tilde{C}_1, \tilde{C}_2 \mid \text{data}) d\mu_2(d\mu_2)$. Evaluating this expression will require specifying a prior $P(\mu_1, \mu_2, \tilde{C}_1, \tilde{C}_2)$. In particular, in large MAS's, one would probably want a prior that biases $\mu_1 - \mu_2$ to have most of its components close to 0.

To illustrate this we consider the case where \tilde{C}_1 and \tilde{C}_2 are known or in some other way fixed (e.g., they're set to their maximum likelihood estimates), and $P(\mu_1, \mu_2 \mid \text{data})$ is well-approximated by a diagonal Gaussian with mean \bar{z} and variances $\bar{\sigma}_i$.⁴ Under these conditions, $\nabla_{\vec{b}} E(G \mid \vec{b}, m, \text{data}) = \int e^{-\sum_i (z_i - \bar{z}_i)^2 / 2\bar{\sigma}_i^2} \nabla_{\vec{b}} \epsilon(\vec{b}) / [(2\pi)^{N/2} \prod_i |\bar{\sigma}_i|] d\vec{z}$, where $\nabla_{\vec{b}} \epsilon(\vec{b})$ is evaluated at $\mu_1 - \mu_2 = \vec{z}$. Since that gradient is itself a Gaussian in \vec{z} times a monomial in \vec{z} , $\nabla_{\vec{b}} E(G \mid \vec{b}, m, \text{data})$ is a Gaussian integral. One can carry through this integral to get a closed form expression, which can then be used in a gradient ascent to find the maxima of $E(G \mid \vec{b}, m, \text{data})$.

The functional form of this closed form expression for the gradient is not very illuminating however. As a pedagogical alternative, we assume that our Gaussian $P(\vec{z} \mid \text{data})$ is sharply peaked, and approximate $\epsilon(\vec{b})$ to second order in \vec{z} about \bar{z} , the peak of our Gaussian:

$$\epsilon(\vec{b}, \vec{z}) \approx \epsilon(\vec{b}, \bar{z}) + (\vec{z} - \bar{z}) \cdot \nabla_{\vec{z}} \epsilon(\vec{b}, \vec{z}) \Big|_{\vec{z} = \bar{z}} + \frac{1}{2} \sum_{i,j} (\vec{z}_j - \bar{z}_j)(\vec{z}_i - \bar{z}_i) \frac{\partial^2}{\partial z_i \partial z_j} \epsilon(\vec{b}, \vec{z}) \Big|_{\vec{z} = \bar{z}}.$$

The second term integrated against our Gaussian $P(\vec{z} \mid \text{data})$ equals 0, since that term has odd symmetry. The first term just contributes $\epsilon(\vec{b}, \bar{z})$ after that integration. Doing the double differentiation in the third term and evaluating at $\vec{z} = \bar{z}$ produces a Gaussian in \vec{z} . We must evaluate the integral over \vec{z} of the product of that Gaussian with $P(\vec{z} \mid \text{data})$. The result is the following approximation:

ing components of \vec{b} to 0 will not affect the numerator of the argument of the erf occurring in Thm. 4, but it will decrease the denominator. Accordingly, our doing this will increase the erf term in Thm. 4 if $\vec{b} \cdot (\vec{\mu}_2 - \vec{\mu}_1)$ is positive, and will decrease it if $\vec{b} \cdot (\vec{\mu}_2 - \vec{\mu}_1)$ is negative. So if $\tilde{D} = \tilde{0}$, the \vec{b} maximizing $\epsilon(\vec{b})$ has the value 0 for all components for which $\vec{\mu}_2 - \vec{\mu}_1$ equals 0. Intuitively, the optimal student ignores all components of the $\vec{y}_A(t)$ for which both actions have the same expected payoff, since those components just contribute an overall noise to the reward signal.

In practice, as discussed in the experiments section below, we can exploit this effect by setting to 0 all components of \vec{b} whose magnitude falls below some preset threshold. Examination of Thm. 4 suggests that we can go further and approximate such a zeroing operation in a parallel fashion. For example, we could set to 0 all components i such that $\left| \frac{(\vec{\mu}_2 - \vec{\mu}_1)_i}{\sqrt{(\tilde{C}_2 + \tilde{C}_1)_{ii}}} \right|$ is small enough. If the random variable i has a computational device associated with it (e.g., if it is a student in a MAS), then by only examining data generated by random variable i during the observation phase in response to the student's actions, the computational device associated with random variable i can determine whether to zero out the associated component of \vec{b} . Then the data associated with that variable need only be communicated to the teacher if the associated component of \vec{b} has not been zeroed out. As discussed in the introduction, this would potentially reduce significantly the computational and communication burdens on the teacher.

In general, even when $\tilde{D} = \tilde{0}$, so that both $G(\cdot)$ and $g(\cdot)$ are linear functions of \vec{y}_A , the \vec{b} maximizing $\epsilon(\vec{b})$ will not equal \vec{c} . In other words, even if both $G(\cdot)$ and $g(\cdot)$ are linear functions of \vec{y}_A , we will not want to have $g(\vec{y}_A) = G(\vec{y}_A)$. This demonstrates that even in this simple scenario, we will want to distort the reward to achieve best performance of the student. The details of how to estimate that optimizing distortion from a finite set of data are discussed in the next section.

To address the general case we must evaluate the expected value of G as a function of \vec{b} . As mentioned just after its presentation, Thm. 2 applies to our special case of normally distributed \hat{y}_1 and \hat{y}_2 and linear $g(\cdot)$, since in this case $P(\delta_m)$ must be a Gaussian. For pedagogical value, the evaluation of the terms in that theorem is performed in Appendix C. The result is as follows:

Theorem 4: $\varepsilon(\vec{b}) \equiv E(G \mid \vec{b}, \vec{c}, \tilde{D}, \mu_2, \mu_1, \tilde{C}_2, \tilde{C}_1, m) =$

$$\begin{aligned} & \frac{1}{2} [\vec{c} \cdot (\mu_2 + \mu_1) + \mu_2 \cdot \tilde{D} \cdot \mu_2 + \mu_1 \cdot \tilde{D} \cdot \mu_1 + \text{Tr}(\tilde{D} (\tilde{C}_2 + \tilde{C}_1))] \\ & + \\ & \frac{1}{2} [\vec{c} \cdot (\mu_2 - \mu_1) + \mu_2 \cdot \tilde{D} \cdot \mu_2 - \mu_1 \cdot \tilde{D} \cdot \mu_1 + \text{Tr}(\tilde{D} (\tilde{C}_2 - \tilde{C}_1))] \times \\ & \text{erf} \left(\sqrt{\frac{m}{2}} \left[\frac{\vec{b} \cdot (\mu_2 - \mu_1)}{\sqrt{\vec{b} \cdot (\tilde{C}_2 + \tilde{C}_1) \cdot \vec{b}}} \right] \right). \end{aligned}$$

Our expected G is maximized by the \vec{b} for which $\nabla_{\vec{b}} \left[\frac{\vec{b} \cdot (\mu_2 - \mu_1)}{\sqrt{\vec{b} \cdot (\tilde{C}_2 + \tilde{C}_1) \cdot \vec{b}}} \right] = \vec{0}$. By examining Thm. 4 we see that which of the zeroes of this quantity we want will depend on whether we want the maximum or the minimum of the quantity $\text{erf} \left(\sqrt{\frac{m}{2}} \left[\frac{\vec{b} \cdot (\mu_2 - \mu_1)}{\sqrt{\vec{b} \cdot (\tilde{C}_2 + \tilde{C}_1) \cdot \vec{b}}} \right] \right)$ occurring in Thm. 4. In turn, which of those we want will depend on the sign of the multiplicative factor $[\vec{c} \cdot (\mu_2 - \mu_1) + \mu_2 \cdot \tilde{D} \cdot \mu_2 - \mu_1 \cdot \tilde{D} \cdot \mu_1 + \text{Tr}(\tilde{D} (\tilde{C}_2 + \tilde{C}_1))]$. So for example, if that sign is positive, then we want the maximum of $\text{erf} \left(\sqrt{\frac{m}{2}} \left[\frac{\vec{b} \cdot (\mu_2 - \mu_1)}{\sqrt{\vec{b} \cdot (\tilde{C}_2 + \tilde{C}_1) \cdot \vec{b}}} \right] \right)$, which occurs when its argument is maximal. Conversely, if the sign of the factor is negative, we want to minimize that argument. In particular, if $\tilde{D} = \vec{0}$, then we want $\vec{b} \cdot (\mu_2 - \mu_1)$ and $\vec{c} \cdot (\mu_2 - \mu_1)$ to have the same sign. In other words, we want \vec{b} and \vec{c} to have the same projection onto the difference in expected \hat{y} 's, $(\mu_2 - \mu_1)$.

Now consider the case where some components of $\mu_2 - \mu_1$ equal 0. Setting the correspond-

can update $g(\cdot)$ (which in this case means updating \vec{b}), then the teacher's goal is use the data it gleans during the observation phase to calculate the optimal \vec{b} , which it then transmits to the student, and which the student subsequently uses to evaluate its reward signals during the teaching phase.

For $\tilde{D} = \tilde{0}$, $\vec{c} = \{1, 1, 1, \dots, 1\}$, and both \tilde{C}_A diagonal, there are two extremal cases. In the first one, $\vec{\mu}_2 - \vec{\mu}_1 = \{1, 0, 0, \dots, 0\}$. In this case, having any of the components of $b_{i \geq 2}$ nonzero will simply result in noise being added to $g(\vec{y}_1(t)) - g(\vec{y}_2(t))$, in the sense that the associated components of $\vec{y}_2(t)$ and $\vec{y}_1(t)$ convey no information about which action is preferable, and therefore can only serve to “confuse” the student's algorithm, $A(M+m+1) = 3/2 + \text{sgn}[\sum_{M+1 \leq i \leq M+m} g(\vec{y}_1(t)) - g(\vec{y}_2(t))] / 2$. (In fact, it is hard to imagine a non-pathological algorithm for which allowing any of the components of $\vec{y}_2(t)$ and $\vec{y}_1(t)$ beyond the first to contribute to the associated rewards can do anything other than decrease performance for this case.) Accordingly, the optimal \vec{b} for this case is $(1, 0, 0, \dots, 0)$.

Intuitively, in the language of sampling theory statistics, having $\vec{b} = (1, 0, 0, \dots, 0)$ rather than $\vec{b} \propto \vec{c}$ introduces bias into the student's algorithm, but more than compensates for that by decreasing the variance so that the total sum of bias and variance - which gives the expected performance - improves [15]. Another way to understand the usefulness of having \vec{b} and \vec{c} not be parallel is to view the student as running a search algorithm to try to find the optimal action. In this perspective, the student is repeatedly sampling the (noisy) surface that maps actions to rewards, with the desire of finding its maximum. Distorting the reward signal then corresponds to modifying a surface to make it easier to search while leaving its maximum intact.

Conversely, consider having $\vec{\mu}_2 - \vec{\mu}_1 = \{1, 1, 1, \dots, 1\}$ and $\tilde{C}_2 = \tilde{C}_1$, so that each sample of $\vec{y}_2(t) - \vec{y}_1(t)$ is an N -fold IID sample of the same underlying Gaussian distribution having mean 1. In this case, $\vec{b} = \{1/N, 1/N, \dots, 1/N\} \propto \vec{c}$ results in $g(\vec{y}_1(t)) - g(\vec{y}_2(t))$ being an average of N IID samplings of the same underlying distribution. Such an average will have smaller variance than would having $\vec{b} = (1, 0, 0, \dots, 0)$, while having the same bias (namely 0). In this case, the optimal \vec{b} is any vector proportional to \vec{c} .

ii) If the random variables $G(\hat{y}_0)$ and $G(\hat{y}_1)$ are symmetrically distributed about their means, then

$$\rho_G(2, m) \leq \min \left[\frac{1}{2}, \frac{\hat{\sigma}_{G,1}^2 + \hat{\sigma}_{G,2}^2}{2m(\hat{G}_2 - \hat{G}_1)^2} \right].$$

Similar bounds hold for $\rho_{\{g_i\}}(2, m)$, with $\hat{\sigma}_{G,A}$ replaced by $\hat{\sigma}_A$ and replaced by \hat{g}_A .

As a particular example of Thm. 3, for normally distributed \hat{y}_0 and \hat{y}_1 , and $G(\cdot)$ that is a linear function of its argument, bound (ii) applies, and we know that modifying the reward signal cannot gain us a factor greater than 2 in normalized performance.

3. DEFINITION OF OUR SPECIAL CASE

We now investigate a particular instance of this general phenomenon for $K = 2$, and in particular what is involved in approaching the performance improvement theoretically allowed according to Thm. 3. For simplicity, we take each of the two distributions $P(\hat{y}_A)$ to be a Gaussian, centered on $\vec{\mu}_2 \equiv (\mu_{1,2}, \dots, \mu_{N,2})$ and $\vec{\mu}_1 \equiv (\mu_{1,1}, \dots, \mu_{N,1})$ respectively, and with (positive definite) covariance matrices \tilde{C}_2 and \tilde{C}_1 , respectively. Also for simplicity, we take $G(\hat{y}_A) = \vec{c} \cdot \hat{y}_A + \hat{y}_A \cdot \tilde{D} \cdot \hat{y}_A$ for some matrix \tilde{D} and vector \vec{c} , and $g(\hat{y}_A) = \vec{b} \cdot \hat{y}_A$ for some vector \vec{b} . Note that the magnitude of \vec{b} will have no effect on the student's decision of which action to take.

Having non-zero \tilde{D} means that we are using a linear reward signal even though we know G is nonlinear. There are several situations that this mismatch is meant to model. Perhaps the most important is where due to computational limitations, the student can only use reward signals that are linear functions of the $\{\hat{y}_1(t)\}$ and $\{\hat{y}_2(t)\}$. In particular, it may be that due to communication restrictions only a subset of the components of the $\{\hat{y}_1(t)\}$ and $\{\hat{y}_2(t)\}$ are transmitted to the student (namely those components i for which b_i is not close to 0), and due to computational restrictions the student can only evaluate linear combinations of those transmitted components to get its reward signal. If there are also computational restrictions on the teacher, restricting how often it

Note that for any values of μ and σ , $\min(\frac{1}{2}, \frac{\sigma^2}{2\mu^2}) \leq \frac{\sigma^2}{\mu^2 + \sigma^2}$; reasonably, the bound in (ii), based on extra restrictions on $P(x)$, never exceeds the bound in (i).

For $m = 1$, we can simplify our notation and write the expected value of G given the naive student and the naive reward as $P(A = 1) \hat{G}_1 + P(A = 2) \hat{G}_2$. $P(A = 1)$ is just the probability that $G(\hat{y}_2) - G(\hat{y}_1) < 0$. Defining the random variable z to be $G(\hat{y}_2) - G(\hat{y}_1)$, $E(z) = \hat{G}_2 - \hat{G}_1 > 0$, and we see that $P(A = 1)$ is just the probability bounded in Lemma (1), with the x in that lemma set equal to z . ($P(A = 2)$ is just 1 minus this probability.)

To evaluate this bound on $P(A = 1)$ we need the variance of z , σ_z^2 . This is just the sum of the variances of $G(\hat{y}_1)$ and $G(\hat{y}_2)$, $\hat{\sigma}_{G,1}^2$ and $\hat{\sigma}_{G,2}^2$, respectively. Accordingly, the bound in Lemma 1(i) is $\frac{\hat{\sigma}_{G,1}^2 + \hat{\sigma}_{G,2}^2}{\hat{\sigma}_{G,1}^2 + \hat{\sigma}_{G,2}^2 + (\hat{G}_2 - \hat{G}_1)^2}$. Together with the facts that $E(G \mid \text{best possible reward and student}) = \hat{G}_2$ and $E(G \mid \text{worst possible reward and student}) = \hat{G}_1$, this means that $\rho_G(2, 1) \leq \frac{\hat{\sigma}_{G,1}^2 + \hat{\sigma}_{G,2}^2}{\hat{\sigma}_{G,1}^2 + \hat{\sigma}_{G,2}^2 + (\hat{G}_2 - \hat{G}_1)^2}$. If we know that $G(\hat{y}_1)$ and $G(\hat{y}_2)$ are symmetric about their means, then we can instead use the tighter of the two bounds in Lemma 1, $\min\left(\frac{1}{2}, \frac{\hat{\sigma}_{G,1}^2 + \hat{\sigma}_{G,2}^2}{2(\hat{G}_2 - \hat{G}_1)^2}\right)$. So if the random variables $G(\hat{y}_2)$ and $G(\hat{y}_1)$ are symmetrically distributed about their means, then we can instead write $\rho(2, 1) \leq \min\left(\frac{1}{2}, \frac{\hat{\sigma}_{G,1}^2 + \hat{\sigma}_{G,2}^2}{2(\hat{G}_2 - \hat{G}_1)^2}\right)$.

For $m > 1$, we need only use the fact that the distribution of the average of m IID samples of any random variable t has the same mean as t and $(1/m)$ times its variance. This gives the following general results:

Theorem 3:

$$i) \quad \rho_G(2, m) \leq \frac{\hat{\sigma}_{G,1}^2 + \hat{\sigma}_{G,2}^2}{\hat{\sigma}_{G,1}^2 + \hat{\sigma}_{G,2}^2 + m(\hat{G}_2 - \hat{G}_1)^2};$$

suggestion that in general we want to have $C(\hat{g}_2 - \hat{g}_1, (\hat{\sigma}_1^2 + \hat{\sigma}_2^2) / m)$ be small if $\hat{G}_2 - \hat{G}_1$ is positive, and large otherwise. Accordingly, $\hat{g}_2 - \hat{g}_1$ should have the same sign as $\hat{G}_2 - \hat{G}_1$. (This is called having a “factored” reward in [14].) Subject to that restriction we should choose the $\{g_i(\cdot)\}$ so as to maximize the difference of $C(\hat{g}_2 - \hat{g}_1, (\hat{\sigma}_1^2 + \hat{\sigma}_2^2) / m)$ from $1/2$. This means minimizing the variances $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$, while having the magnitude of $\hat{g}_2 - \hat{g}_1$ be as large as possible. Intuitively, this means that our signal will be maximally visible compared to our noise. (This is called having “high learnability” in [14].)

Taken as a whole, we can use these considerations to advise us on what kind of functional $U(\{g_i(\cdot)\}, \text{data})$ we should maximize to set the $\{g_i(\cdot)\}$. For example, it makes sense to have that functional depend on the posterior probability that $\hat{g}_2 - \hat{g}_1$ has the same sign as $\hat{G}_2 - \hat{G}_1$, π , and on data-based estimates of $|\hat{g}_2 - \hat{g}_1|$, of $\hat{\sigma}_1^2$, and of $\hat{\sigma}_2^2$. It should be an increasing function of its estimate of $[m |\hat{g}_2 - \hat{g}_1| / (\hat{\sigma}_1^2 + \hat{\sigma}_2^2)]$ and of π . Since it is a posterior probability, π automatically has a bias preferring that all the $\{g_i(\cdot)\}$ equal $G(\cdot)$, a bias arising from its prior. To ensure that maximizing $U(\{g_i(\cdot)\}, \text{data})$ is unlikely to result in a performance degradation compared to the naive student and reward, one may wish to augment that bias by having $U(\{g_i(\cdot)\}, \text{data})$ weight the π term more heavily as m shrinks. Intuitively, in keeping with Thm. 2, one would expect that choosing the $\{g_i(\cdot)\}$ that maximize any reasonable $U(\{g_i(\cdot)\}, \text{data})$ that has these characteristics should result in good performance.

For $K = 2$ it is straight-forward to bound $\rho_G(K, m)$ (and therefore the potential performance improvement entailed by distorting the reward function) independently of the details of our distributions. To do so we will use the following “one-sided Chebychev’s inequalities”, derived in Appendix B:

Lemma 1: For any real-valued random variable x with mean μ and variance σ^2 ,

$$\text{i) } P(\text{sgn}(x) \neq \text{sgn}(\mu)) \leq \frac{\sigma^2}{\mu^2 + \sigma^2},$$

and if x is symmetric about μ , then

$$\text{ii) } P(\text{sgn}(x) \neq \text{sgn}(\mu)) \leq \min\left(\frac{1}{2}, \frac{\sigma^2}{2\mu^2}\right).$$

over $\mathfrak{R}^N \setminus R$, there is a non-zero probability that the empirical mean of our sample of $g(\hat{y}_K)$ exceeds those of all of the samples of the $g(\hat{y}_{i < K})$ even as $K \rightarrow \infty$. In such a situation, $\rho_g(K, m)$ is less than 1 even as $K \rightarrow \infty$. Accordingly, if we can identify such “forbidden regions” R and then incorporate them into $g(\cdot)$, we will have drastically improvement performance (compared to using the naive reward) for the case of an infinite number of actions.

On the opposite side of the spectrum, we will often be interested in the special case where $K = 2$ and our distributions can be parameterized by their means and variances. To address this case, define δ_m as the random variable given by the average of m IID samples of $g_2(\hat{y}_2) - g_1(\hat{y}_1)$. Then $P(A = 2 \mid g_1, g_2, m)$ is just the probability that δ_m is positive. Similarly, $P(A = 1 \mid g_1, g_2, m)$ is just the probability that δ_m is negative. So $E(G \mid g_1, g_2, m) = \hat{G}_1 \Pr(\delta_m \leq 0) + \hat{G}_2 \Pr(\delta_m > 0)$.

Moreover, by direct expansion, $E(\delta_1) = \hat{g}_2 - \hat{g}_1$, and the variance of δ_1 equals $\hat{\sigma}_1^2 + \hat{\sigma}_2^2$. Iterating this rule for summing pairs of random variables, $E(m\delta_m) = m(\hat{g}_2 - \hat{g}_1)$ and the variance of $m\delta_m$ equals $m(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)$. Accordingly, $E(\delta_m) = \hat{g}_2 - \hat{g}_1$, and the variance of δ_m equals $(\hat{\sigma}_1^2 + \hat{\sigma}_2^2) / m$. This establishes the following result:

Theorem 2: Assume that the cumulative distribution function of δ_m is uniquely specified by its mean and variance. Let $C(E(\delta_m), \sigma^2(\delta_m))$ be that cumulative distribution function evaluated at 0. Then

$$E(G \mid g_1, g_2, m) = \hat{G}_1 + [\hat{G}_2 - \hat{G}_1] [1 - C(\hat{g}_2 - \hat{g}_1, (\hat{\sigma}_1^2 + \hat{\sigma}_2^2) / m)].$$

So for instance Thm. 2 applies if δ_m is normally distributed. As a particular example, investigated in detail below, the assumption in Thm. 2 holds when \hat{y}_0 and \hat{y}_1 are both normally distributed and both $g_1(\cdot)$ and $g_2(\cdot)$ are linear functions of their arguments. This is because under those circumstances δ_1 is a gaussian random variable, and therefore so is δ_m .

More generally, Thm. 2 provides guidance on how to set the $\{g_i(\cdot)\}$ even when it does not strictly apply, for example when we do not have strong prior beliefs concerning how the $\{\hat{y}_i\}$ are distributed, and/or the student does not use the naive algorithm. Thm. 2 makes the very reasonable

$E(G \mid \text{naive student}, \{g_i\}, K, m) = \sum_{A=1}^K [\hat{G}_A \times P(\text{naive student chooses action } A \mid \{g_i\}, K, m)]$,
the following result is derived in Appendix A:

$$\textbf{Theorem 1: } \rho_{\{g_i\}}(K, m) = \frac{\sum_{A < K} \left[(\hat{G}_K - \hat{G}_A) \int P_{A, g_A, m}(t) \prod_{j \neq A} C_{j, g_j, m}(t) dt \right]}{\hat{G}_K - \hat{G}_1}.$$

To illustrate this theorem, consider the case of the naive reward. Intuitively, the more actions the student can take, the more action-reward data sets it will look at at the end of the teaching phase. Any one of those data sets may, by “statistical fluke”, have higher empirical mean than the sample corresponding to the optimal action. So the more actions the student can take, the more likely it is to find *some* action which appears to be better than that of what is in fact the optimal action. Thus, the more actions it can take, the more likely the student is to not choose the optimal action.

By using Thm. 1, we can illustrate this phenomenon for the case where all the variables $\mathfrak{y}_{i < N}$ are identically distributed, and therefore so are the $G(\mathfrak{y}_{i < N})$. By that theorem, we can write $\rho_G(K, m) = (K - 1) \int P_{1, G, m}(t) [C_{1, G, m}(t)]^{K-2} C_{K, G, m}(t) dt$. In turn, we can write this as $1 - \int P_{K, G, m}(t) [C_{1, G, m}(t)]^{K-1} dt$, since $\int \frac{d}{dt} (\prod_i C_{i, G, m}(t)) dt = 1$. Consider the common case where the support of $P_{1, G, m}(t)$ is infinite. In this case, in the limit of large K , the product of cumulative distributions in our integrand goes to 0 everywhere away from infinity. Accordingly, so long as $P_{K, G, m}(t)$ is nowhere singular, the integral goes to 0, and $\rho_G(K, m)$ reduces to 1. Thus, for fixed m , in the limit of a large number of possible actions, the naive student and reward will do as *poorly* as possible.

For the same problem not to befall the use of the distorted rewards $\{g_i(\cdot)\}$, even if the $g_{i < K}(\cdot)$ are identical, it is necessary that $C_{1, g_1, m}(t)$ reaches 1 before $C_{K, g_K, m}(t)$ does. If both $P(\mathfrak{y}_1)$ and $P(\mathfrak{y}_K)$ are nowhere zero, this in turn requires that $g_1(\cdot) \neq g_K(\cdot)$. However there are cases where all the $g_A(\cdot)$ are identical, but still $\rho \neq 1$. In particular, consider the case where there exists a region $R \subset \mathfrak{R}^N$ across which \mathfrak{y}_1 is forbidden, while \mathfrak{y}_K can occur there with non-zero probability. Then by having $g_1(\cdot) = g_K(\cdot) \equiv g(\cdot)$ and giving values to all $g(\mathfrak{y} \in R)$ that are larger than any of the values

$G(\cdot)$ for all A the *naive reward*. For the special case of $K = 2$, we can write the naive student's algorithm as $A(M+m+1) = 3/2 + \text{sgn}[\sum_{M+1 \leq t \leq M+m} g_2(\hat{y}_2(t)) - g_1(\hat{y}_1(t))] / 2$, where $\text{sgn}(z \in \mathfrak{R})$ is defined to be the sign of z when z is nonzero, and to be zero otherwise.

To normalize how much of an improvement we can possibly get by distorting the reward signal, we define a fractional improvement in performance:

$$\rho_{\{g_i\}}(K, m) \equiv \frac{E(G \mid \text{best possible reward and student}) - E(G \mid \text{rewards } \{g_i(\cdot)\} \text{ and naive student})}{E(G \mid \text{best possible reward and student}) - E(G \mid \text{worst possible reward and student})}.$$

$\rho_G(K, m)$ is defined as $\rho_{\{g_i\}}(K, m)$ when the naive rewards are used, so all the $g_i(\cdot)$ equal G .

$\rho_{\{g_i\}}(K, m)$ is implicitly a function of the distributions governing the \hat{y}_A and the choice of the $\{g_i(\cdot)\}$. $\rho_G(K, m)$ also depends on those distributions, in addition to depending on $G(\cdot)$. $\rho_G(K, m)$ is a normalized measure of the largest performance improvement potentially achievable by distorting the reward functions and/or the student. $\rho_G(K, m) - \rho_{\{g_i\}}(K, m)$ measures the actual normalized performance improvement if the naive student is used with the reward functions $\{g_i(\cdot)\}$ rather than if it is used with the single reward function $G(\cdot)$ for all actions.

For the analysis below it will be useful to define $P_{A,f,m}(x \in \mathfrak{R})$ for an arbitrary function $f(\hat{y}_A)$ as the probability density function over the average of m IID samples of $f(\hat{y}_A)$.³ Define the K means $\hat{g}_A \equiv E(g_A(\hat{y}_A))$, and define the associated variances $\hat{\sigma}_A^2 \equiv E([g_A(\hat{y}_A) - \hat{g}_A]^2)$. Also define the K cumulative distribution functions as the integrals of the $\{P_{A,g_A,m}\}$: $C_{A,g_A,m}(t \in \mathfrak{R}) \equiv \Pr(\hat{y}_A : \text{the average of } m \text{ IID samples of } g_A(\hat{y}_A) \leq t)$. Finally, define $\hat{G}_A \equiv E(G(\hat{y}_A))$ and $\hat{\sigma}_{G,A}^2 \equiv E([G(\hat{y}_A) - \hat{G}_A]^2)$, and label the K actions in order of ascending \hat{G}_A , from 1 to K .

2. GENERAL FORMULAS FOR PERFORMANCE IMPROVEMENT

In general, to calculate $\rho_{\{g_i\}}(K, m)$ we need only calculate $E(G \mid \text{naive student}, \{g_i\}, K, m)$, since the other three terms in the definition of $\rho(\cdot, \cdot)$ are all either \hat{G}_K or \hat{G}_1 . Using the expansion

time $t \in \{M + 1, \dots, M + m\}$ the student takes all such actions, in succession. There are a set of N real-valued random variables which are sampled once after each such action. This generates K separate N -dimensional vectors, $\{\hat{y}_i(t) \equiv (y_{1,i}(t), \dots, y_{N,i}(t)) : i \in \{1, \dots, K\}, t \in \{M + 1, \dots, M + m\}\}$. In each such sampling, the distribution governing the N -dimensional random variable \hat{y}_i is determined solely by the student's associated action, and in particular does not depend on the results of any other samplings. Furthermore, the rule relating the student's action and the distribution over the N variables do not change in time. So another way to view a set of m K -tuples of samplings is as the generation of K distinct sets of m independent and identically distributed (IID) samplings, one set for each of the K separate N -dimensional random variables associated with each of the student's K possible actions.

At each moment $t \in \{M+1, \dots, M+m\}$, in response to each of its K actions, the student will receive K associated reward signals, with values $\{g_i(\hat{y}_i(t))\}$ for some functions $\{g_i(\cdot)\}$ (i ranges over the possible actions). The student will then at time $M+m+1$ use that set of Km reward signals to estimate which of its K possible actions A will, if used forevermore, likely result in the highest possible value of $\sum_{t \geq M+m+1} g_A(\hat{y}_A(t))$. Due to the IID nature of the generation of the samples, this is equivalent to estimating which action A will result in the highest possible value of $g_A(\hat{y}_A(M+m+1))$.

The student's utility is $\sum_t G(\hat{y}_{A(t)}(t))$ for some function $G(\cdot)$, where $A(t)$ is its action at time t . We will judge the student's performance (and therefore the choice of the $\{g_A(\cdot)\}$) not in terms of $\sum_{t \geq M+m+1} g_A(\hat{y}_A(t))$, but in terms of the true utility, $\sum_{t \geq M+m+1} G(\hat{y}_{A(t)}(t))$, i.e., in terms of $G(\hat{y}_{A(M+m+1)}(M+m+1))$. So given a particular specification of the student's estimator for predicting which A to use, our goal is to choose the set of reward functions $\{g_i(\cdot)\}$ that optimizes the value of $G(\hat{y}_{A(M+m+1)}(M+m+1))$ that will result from the student's using that estimator.

For simplicity, we assume the student performs its estimation of which action to use with the maximum likelihood unbiased estimator of the means of the K distributions governing the generation of the $\{\hat{y}_A\}$.¹ In other words, we assume the student picks the action given by $\text{argmax}_A [\sum_{M+1 \leq t \leq M+m} g_A(\hat{y}_A(t))]$.² We call this the *naive student*, and we call having $g_A(\cdot) =$

ations one can often partially “parallelize the teacher”, by distributing to those computational devices an approximate version of the calculation of whether the associated component of the linear filter is low enough to be set to zero. Under this approach, during the observation phase, each of the computational devices collects the data associated with its random variable. Then at the end of the observation phase, all the devices look at their accumulated data, and only those devices that decide that they should *not* zero the associated component communicate their data to a central teacher. Then that teacher performs its calculation of the best linear filter, but only considering those components of the filter whose associated devices it has data from. In this scheme both the computational and communication burdens on the teacher may be reduced substantially, compared to having the teacher receive and process all the data from all the random variables. This is in addition to the reduction in such burdens already enjoyed by the students, via their having their rewards determined by the teacher.

Sections 3 and 4 are more intuitive than Section 2. As much as possible, we have written those two sections so that the reader can skip to them directly after having read Section 1.

The results of this paper demonstrate that there are scenarios in which an appropriate choice of reward signal can result in an extremely large improvement in the performance of the student. They also show that it is possible, at least in a simple scenario, to exploit this phenomenon by having a teacher observe a system, and based on that observation, tailor the reward signal the student receives. Doing this markedly improves the student’s subsequent performance, all at little cost both in extra communication overhead on the system as a whole and in computational overhead on the student.

1. GENERAL PROBLEM DEFINITION

We consider one of the simplest possible RL scenarios. There is a single teacher and a single RL-based student. The student can only take one of K possible actions, $A \in \{1, \dots, K\}$. At each

distribution-independent bounds on the maximal gain in performance potentially achievable by distorting the reward function for the special case where $K = 2$. The results of this section hold for arbitrary reward functions, including functions that are not linear filters and/or that depend on the student's action.

In Section 3 we present a preliminary investigation of how some of the potential improvement calculated in Section 2 might actually be realized. We do this by analyzing in detail a special, simple version of the case where the reward function is given by a linear filter that is the same for any of the student's action. For this special case we calculate in closed form the energy function mapping the filter to the associated expected value of the student's performance at time $M + m + 1$. This function is parameterized by the distributions governing the relationship between the student's actions and the rest of the system. During its observation phase the teacher can form a Bayesian posterior over those distributions, and therefore a posterior expected energy function. In Section 4 we present an approximate calculation of the posterior expected energy function. We then present a simple-minded gradient ascent scheme for that posterior expected energy function that the teacher can use to (try to) calculate the optimal filter.

In Section 5 we present the results of experimental tests of that scheme. In particular, we investigate an approximation to the gradient ascent in which the smallest of the components of the filter are set to 0 after the ascent has completed. With this approximation, the communication overhead in generating the student's reward signal at each moment in the teaching phase is minimal, an important consideration in real MAS's. Indeed, in many MAS's each student can only poll a small number of random variables at each time step. For such a system, having a teacher determine which variables the student should poll (e.g., by setting many components of a linear filter to 0) is more a necessity than a luxury. This approximation also reduces the computational overhead on the student, another important practical concern.

A final advantage of this approximation arises when there are many random variables in the student's environment that affect its true utility, and many of those variables are accompanied by computational devices (e.g., if those random variables are other students in a MAS). In such situ-

with balancing exploration vs. exploitation, and the like. In addition, for a large system with many random variables, the student may have difficulty discerning the “echo” of its actions in the values $G(\zeta_t)$, since the effect of those actions could be swamped by all the other processes in the system. In essence the student faces a signal/noise problem. However, due to its superior observational abilities, computational resources, and prior knowledge, the teacher can more directly discern the effects of the student’s actions. It can then distort the rewards received by the student to reflect this deeper understanding of the system. For example, the teacher can accentuate the contribution to the reward coming from those random variables that depend strongly on the student’s actions, in a fashion reminiscent of Kalman filters. In essence, in this setting the teacher is trying to “compress” a sophisticated analysis of all the relevant information it has access to into a form usable by the computationally restricted student (namely, into a reward function). It then transfers that information to the student, and in this way shoulders much of the student’s computational burden.

This paper investigates this issue of how the teacher should set the rewards to improve signal/noise for some very simple (and therefore tractable in closed form) scenarios. In these scenarios, the teacher first observes the overall system for some “observation phase”, $t \in \{1, \dots, M\}$. The teacher then uses that data to set a reward function. The teacher has nothing to do with the student subsequent to this calculation. The reward signal received by the student at each moment during a subsequent “teaching phase”, $t \in \{M + 1, \dots, M + m\}$, is given by applying the reward function calculated by the teacher to the state of the full system at each such t . (If it is just a linear filter, calculation of such a reward imposes minimal computational overhead on the entity calculating the rewards, which may be the student itself.) The student then uses those signals to choose what action to take at $t = M + m + 1$. So for example, in the case of online, continual RL, we would have $m = 1$. The algorithm used by the student to make its choice was known ahead of time by the teacher when the teacher was deciding on the filter.

In Section 1 we present our general problem in detail. In Section 2 we derive the formula for the performance (as measured by the student’s true utility) accompanying any particular reward function, as a function of m and of the number of actions K the student can take. We then present

INTRODUCTION

Consider the following scenario:

- 1) There is a “student” running a Reinforcement Learning (RL - [2, 3, 7, 8, 9, 11]) algorithm, who knows relatively little *a priori* concerning the relationship between its observations, its actions, and the responses of the environment.
- 2) There is a separate “teacher” who watches the student as well as the rest of the system, and knows the student’s “true” utility function. The teacher (potentially) knows the form of the probability distributions underlying the full system’s dynamics.
- 3) The teacher determines the rewards that the student receives.

How should the teacher set the student’s rewards to most benefit the student, i.e., to maximize the student’s true utility? This question arises particularly often in RL-based Multi-Agent Systems (MAS’s - [3, 6, 10]). Invariably in such systems different agents have access to different amounts of global information and have different computational resources. Moreover, often there is nothing preventing the more powerful and knowledgeable of the agents from modifying the calculations of the rewards received by the other agents. So there is no *a priori* reason that they cannot play the role of teacher.

One answer to our question would be for the teacher to simply provide the student with the conventional reward signal associated with the student’s true utility function. For example, if time t is discrete and integer-valued, the state of the full system at time t is the Euclidean vector ζ_t , and the student’s goal is to maximize an undiscounted sum of values $G(\zeta_t)$, then the teacher could provide the student the reward signal $G(\zeta_t)$ at each moment t .

An alternative would be for the teacher to use its superior insight to “steer” the student, by distorting the rewards received by the student in such a way as to induce the student to learn more effectively. This could potentially help the student with solving its credit assignment problem,

DISTORTING REWARD FUNCTIONS TO IMPROVE REINFORCEMENT LEARNING

by David H. Wolpert¹, Michael H. New¹, and Ann M. Bell²

1 - NASA Ames Research Center.

2 - Caeulum Research Corporation.

Contact dhw at NASA Ames Research Center, N269-1, Moffett Field, CA 94035, 650-604-3362,
dhw@ptolemy.arc.nasa.gov

Abstract: This paper investigates distorting the reward function to improve the performance of a reinforcement learning algorithm. This issue is particularly important when many random variables contribute to performance, and in particular in large, heterogeneous, multi-agent systems. For tractability, we concentrate on a very simple scenario in which the utility of a “student” is an undiscounted sum of rewards, and each such reward is a sample of a distribution over a multi-dimensional Euclidean space, where the precise distribution sampled at time t is determined by the student’s action at time t . First we derive the formula for the amount of improvement in performance possible by distorting the reward function. We show that as the number of actions the student can take is increased, using *non*-distorted reward functions results in the worst possible performance with probability 1. We then derive some general upper bounds on the amount of possible improvement in performance for the case where the student can only choose between two possible actions. Next we analyze a particular instance of this scenario in which the underlying distributions are Gaussian. We derive exact formulas for how much performance improvement is possible with a particular parameterized class of distortions of the reward function. We then derive a Bayesian algorithm for how a “teacher” should estimate from a finite set of data which of the distortions from such a class to use. We end with computer experiments verifying the gain in performance entailed by use of that algorithm, and discuss the general implications of this work for large, heterogeneous, multi-agent systems.

Keywords: Multi-agent systems, reward functions, optimal teaching, Bayesian learning

Running head: Distorting reward